Audio Classification Model Comparison

Alessandro Di Stefano
Department of Computer Science
University of Milan
Milan, Italy
alessandro.distefano2@studenti.unimi.it

***Abstract*—Audio classification is a well-known problem for which many algorithms have been proposed in the past. However each algorithm usually has different pros and cons when it comes to data preparation, feature extraction and complexity versus accuracy. In this article I want to explore and compare three different models such as KNN, Random Forest and CNN on both binary and multi-class classification tasks. Also, I show different EDA and dimensionality reduction tools to use when accuracy is not the only metric that needs to be optmized as is often the case in a business context where stakeholders want to understand the main features used for the modeling phase.**

I. Introduction

With the rapid success of ML and DL approaches in the last years, many audio Classification methods have been proposed as audio data analysis is still one of the most important research topics with a large number of applications in real life including brainwave entrainment beats [1], Urban sound [2], genre classification [3] and anomalous audio detection [4]. ML and DL methods are usually more flexible and tend to perform better than traditional methods [5]. However, their performance and generalization capabilities can be strongly affected by the feature selection method which represents a mandatory step before the modeling phase [6]. On DL side the problem is more complex: data-hungry models like CNNs cannot often be trained from scratch and as such other solutions must be sought to avoid a lack in the generalization process due to the over-fitting problem. The aim of this paper is to apply different learning models on two different tasks: fake-audio binary classification and genre multi-classification. Concerning ML models, different EDA and dimensionality reduction methods are presented and discussed in depth. This choice is motivated by the fact that stakeholders in any business environment often want to extract some value from the data. Thus EDA process plays an important role in any ML/DL pipeline as from one side it improves the model comprehension while from the other side it can generate value and provide insights. A CNN is also selected as main DL method as it has proved to be very effective when it comes to classify audio [7]. The rest of this paper is organized as follows. Section 2 describes the proposed approach used to perform the sound classification. Section 3 introduces the experimental setup, while Section 4 shows the final results. Finally, section 5 outlines some possible improvements.

II. The proposed approach

The approach includes the following steps:

- Feature extraction
- Feature analysis and exploration (EDA)
- Insight extraction (for business purpose)
- Modeling phase

For some models feature extraction and reduction phases can be managed by the model itself. The impact of each feature strategy has been measured using clustering algorithm (see III-B). Finally for each model an evaluation score has been reported on a validation dataset.

A. Feature extraction

For each frame the following features have been extracted:

- First 20 Mel-Frequency Cepstral Coefficients (MFCCs) which are a representation of the signal where the frequency bands are distributed according to the mel-scale
- Spectral Centroid which is computed as the weighted mean of the frequencies present in the sound
- Zero crossing rate that is the rate at which the signal changes from positive to negative or back
- Spectral bandwidth which represents the difference between the upper and lower frequencies
- Spectral flatness value to differentiate between noise-like and tone-like sounds
- RMSE which measures the average energy of the audio
- Spectral Rolloff (cutoff at 85%)
  For each of the above features mean, std and max values are computed over all frames.

B. EDA: ECDF curve

ECDF curves can be used to display the data points in a sample from lowest to highest against their percentiles. A curve is showed for a specific class and color. Thus different classes can be compared with respect to the ECDF of a specific feature. As such very discriminating features can be discovered using this kind of chart. So one can select the first k features with a greater discriminative power just looking at the distance among different distribution curves. Clearly this is a quality approach, yet it can be very effective. Also a chi2square [8] test can be run to measure the strength of the difference between two distributions.

C. EDA: Spearman Correlation

The Spearman's rank-order correlation is the non-parametric version of the Pearson product-moment correlation which is appropriate when the relationship between variables is not

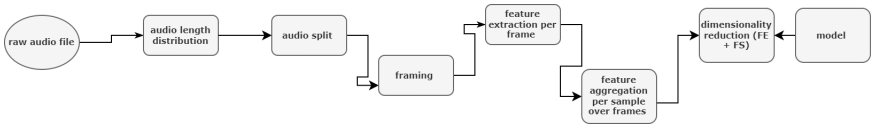

Fig. 1. Block diagram for ML approaches.

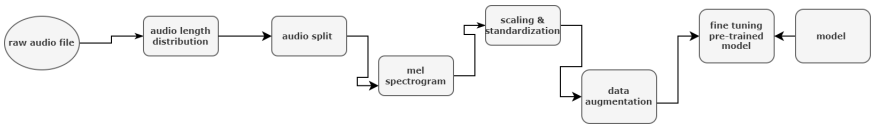

Fig. 2. Block diagram for DL approach.

linear. A value of +1 means a perfect association of rank, a value of 0 means no association of ranks whereas a value of -1 means a perfect negative association between ranks

D. Clustering

For this study K-means is used as clustering algorithm. Even if it considers only convex clusters, it's often used thanks to its simplicity. Clustering is a power data mining tool to discover patterns and groups of homogeneous data. Once all clusters are built one can describe and analyze each of them to make possible hypotheses about the data in a cluster. In addition, these kind of information can be used for stakeholder reporting and engagement. K-means groups are compared against the real labels to measure the quality of the features used.

E. Dimensionality Reduction

A higher number of features makes the model more complex and weaker as the amount of data needed to generalize the model accurately increases exponentially (a problem known as curse of dimensionality) [9]. Feature selection and Feature extraction are two different techniques to reduce the number of variables; the former requires a transformation on data which usually leads to a loss of interpretability; the latter aims to select only the relevant features and is well suited when one wants to have a more readable and interpretable subset of features, yet it can be costly from the computational point of view. For this work, PCA [10] and ISOMAP [11] are used as feature extraction methods while Spearman Coefficient [12], ECDF and Permutation Feature Importance [13] have been selected as feature selection tools. It's worth to note that any CNN can be considered as a natural feature extractor.

F. Modeling

Random forest and KNN models are used as ML approaches for the process pipeline represented in Fig.1 while a pretrained CNN has been adopted as DL approach and its process pipeline is showed in Fig.2. As one can see there is a common part for each pipeline related to the pre-processing of the raw audio signal. Both approaches come with some advantages. For example, ML approaches make easier to understand the most important features as they require a feature selection and extraction phase. On the other hand, DL models can actually perform better in specific domains like image [14] and text classification [15].

III. EXPERIMENTAL SET-UP

A. datasets

To assess the approaches, two classification datasets have been used. The DeepFake [16] dataset is a collection of 8 audio clips from real human speech and 56 ones coming from a generative model which generated some speech from the "real" audio clips. From this dataset a new data sample has been created with each observation lasting 1 second. The GTZAN [17] dataset comprises 10 genres with 100 audio files each, all having a length of 30 seconds. In this case 25 out of 100 of samples per genre are assigned to the validation dataset before being split into new audio files lasting 5 seconds. All data has a sample rate of 22050Hz.

B. feature evaluation

Clustering output is used to evaluate each feature selection/extraction strategy taking into account the true majority label for each cluster. Specifically, a good strategy should lead to a better separation among real classes. Clearly the outcome depends actually from the clustering itself too, but this work wants to point out the importance of selecting the right feature selection/extraction strategy. Also, clustering quality change with respect to a specific strategy has been assessed using silhouette score as a quality metric. [18] In this work K-means is used and the number of clusters will be always equal to the expected ones.

C. ECDF

ECDF curves are used as first way to compare different feature distributions. For both datasets the first 30 features have been selected as input for clustering. The following tables shows the results according to the majority real label and silhouette score for each dataset. Silhouette score is calculated up to $n + 1$ clusters where n is the number of expected clusters. K-means algorithms assumes convex and homogeneous clusters and as such we expect to have a similar deviation from the Silhouette score of the single cluster to the average silhouette score among all clusters.

cluster	num. obs	% real
0	4572	.388
1	2928	.673

TABLE I
K-MEANS OUTCOME USING ECDF FEATURES - DEEPFAKE DATASET

n. clusters	2	3
std Silhouette	.041	.085

TABLE II
CLUSTERING SCORE USING ECDF - DEEPFAKE DATASET

cluster	num. obs	% majority	maj.label
0	476	.81	classical
1	559	.30	blues
2	653	.30	pop
3	628	.29	disco
4	785	.175	reggae
5	618	.483	pop
6	440	.32	blues
7	928	.46	metal
8	432	.5	reggae
9	465	.42	jazz

TABLE III
K-MEANS OUTCOME USING ECDF FEATURES - GENRE DATASET

n. clusters	2	3	4	5	6	7	8	9	10	11
std Silhouette	.0219	.0372	.0804	.0882	.0707	.0813	.0510	.0517	.0509	0479

TABLE IV
CLUSTERING SCORE USING ECDF - GENRE DATASET

From table I it can be observed a good separation between classes using ECDF features as input for the clustering. On the other hand, multi-class classification is far more complex as we can see from table III. In this last case, three classes are missing such as hiphop, country, rock. However classical, metal, reggae and pop clusters have been recognized successfully (almost 50% of correct labels). Finally one can note from table IV a drop in the silhouette deviation after selecting more than 8 clusters as clusters are separated reasonably better

D. Non-linear correlation

Spearman correlation has been calculated and first 30 less-correlated features are selected.

From the table V one can observe outcomes similar to those of the ECDF method. On Genre Dataset three classes are missing such as country, hiphop and rock (same classes as ECDF method). However cluster outcomes seem to be slightly better than ECDF method.

E. Feature reduction - PCA

Unlike previous methods, PCA is actually a compression method which applies a transformation (rotation) on the features. Concerning DeepFake Dataset, 5 components were enough to have a good data approximation, while for Genre Dataset 15 components have been selected.

From the table VIII one can observe outcomes similar to those of the ECDF method. On Genre Dataset (table IX) two classes are missing such as rock and jazz and as such outcomes are slightly better than before.

F. Feature reduction - ISOMAP

Unlike PCA, ISOMAP applies a non-linear transformation. For both datasets, 30 components have been selected.

From the table XII one can observe, regarding Genre Dataset, better results: just 2 classes are missing such as rock and country. Another observation is having a really constant value as silhouette score over the different numbers of classes, this is due to the nonlinearity transformation which in turn makes the K-means not a good fit for this kind of data.

G. Clustering Exploration

After selecting a feature selection/extraction method, one can try to describe clustering output in order to get some insights from the data itself. In case of binary classification, looking at the major difference on a specific feature and then feed this information to a heatmap can be a way to do that. In case of more than two clusters, one can try to analyze the variability with respect to a specific subset of features. For example, figure 5 shows the heatmap for the first 20 features associated with a big variability over all clusters. One can observe many low values associated with the blues class and very high values for the rock one.

H. Modeling

For both datasets I applied the same approach for this phase. Regarding KNN model only the first 20 features are selected from the computed heatmap. In fact, this model is affected by the curse of dimensionality problem above 10-20 dimensions [19]. On random forest side, I first found the best fit using all of the original features whose number

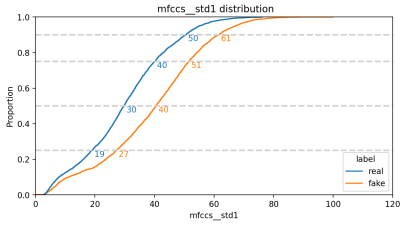

Fig. 3. ECDF of a discriminative feature for the FakeSpeech Dataset

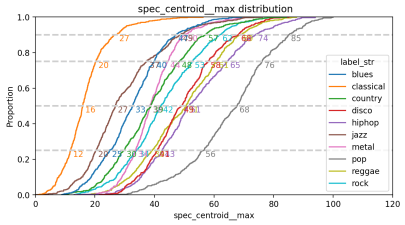

Fig. 4. ECDF of a discriminative feature for the Genre Dataset

cluster	num. obs	% real
0	4761	0.40
1	2739	0.672

TABLE V
K-MEANS OUTCOME USING SPEARMAN FEATURES - DEEPFAKE DATASET

cluster	num. obs	% majority	maj.label
0	204	.78	classical
1	622	.37	reggae
2	1038	.37	metal
3	536	.40	pop
4	393	.36	reggae
5	641	.22	jazz
6	508	.73	classical
7	964	.21	disco
8	561	.34	blues
9	517	.48	pop

TABLE VI
K-MEANS OUTCOME USING SPEARMAN FEATURES - GENRE DATASET

n. clusters	2	3	4	5	6	7	8	9	10	11
std Silhouette	.0505	.0228	.0612	.056	.0523	.057	.052	.0542	.0548	0936

TABLE VII
CLUSTERING SCORE USING SPEARMAN SCORE - GENRE DATASET

cluster	num. obs	% real
0	4616	.39
1	2884	.672

TABLE VIII
K-MEANS OUTCOME USING PCA FEATURES - DEEPFAKE DATASET

cluster	num. obs	% majority	maj.label
0	416	.8	classical
1	694	.233	reggae
2	568	.341	reggae
3	856	.478	metal
4	692	.252	disco
5	401	.531	classical
6	554	.252	country
7	248	.443	hiphop
8	878	.203	blues
9	677	.51	pop

TABLE IX
K-MEANS OUTCOME USING PCA FEATURES - GENRE DATASET

n. clusters	2	3	4	5	6	7	8	9	10	11
std Silhouette	.0132	.0624	.0811	.0883	.082	.0827	.0695	.0494	.0548	.049

TABLE X
CLUSTERING SCORE USING PCA - GENRE DATASET

cluster	num. obs	% real
0	4664	.38
1	2836	.7

TABLE XI
K-MEANS OUTCOME USING ISOMAP FEATURES - DEEPFAKE DATASET

cluster	num. obs	% majority	maj.label
0	664	.84	classical
1	863	.23	blues
2	726	.26	disco
3	482	.44	reggae
4	538	.69	pop
5	776	.25	reggae
6	426	.40	jazz
7	556	.354	hiphop
8	816	.52	metal
9	137	.83	hiphop

TABLE XII
K-MEANS OUTCOME USING ISOMAP FEATURES - GENRE DATASET

n. clusters	2	3	4	5	6	7	8	9	10	11
std Silhouette	.00034	.05554	.06033	.05848	.06512	.06187	.06745	.07046	.06824	.06496

TABLE XIII
CLUSTERING SCORE USING ISOMAP - GENRE DATASET

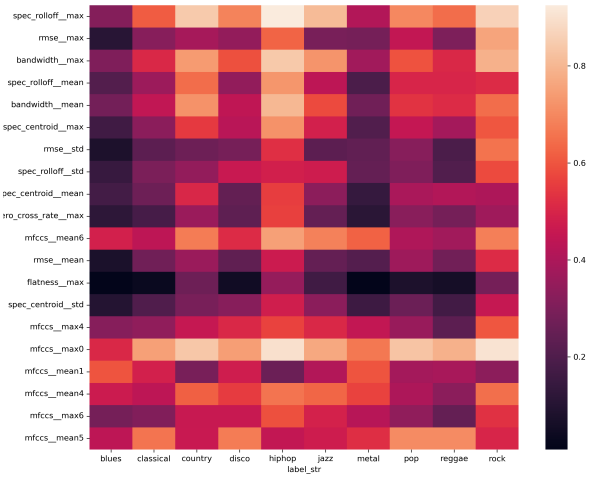

Fig. 5. Heatmap for the genre clusters

Model	KNN	Approx. RF	RF	CNN
F1-Score	.815	.912	.936	.975

TABLE XIV
CAPTION

Genre	KNN	Approx. RF	RF	CNN
blues	.236	.319	.366	0.564
classical	.949	.887	.927	.934
country	.37	.212	.314	.8
disco	.4	.386	.431	.821
hiphop	.146	.188	.171	.79
jazz	.465	.316	.392	.752
metal	.5	.5	.522	.805
pop	.78	.776	.829	.756
reggae	.181	.201	.185	.562
rock	.203	.253	.192	.532
summary	.422	.404	.433	0.73

TABLE XV
CAPTION

is not so high for this kind of model [20], then I selected the best subset of features using the permutation importance method and considering only the first 30 features. Methods like permutation importance can be used wen a data sample is small (so overfitting probability is higher) or one wants to find a smaller subset of features which explains with a reasonable approximation the target variable. Both RF (considering all of the features) and approximate RF (i.e. selecting a subset of features) models are reported. Finally regarding the DL approach, I finetuned a pretrained CNN model that is **resnet34** [21]. For the genre dataset, I added a further penalty to the loss function using the weight decay approach in order to reduce the overfitting phenomenon for this more complex dataset [22]. Also data augmentation approach has been used during the training steps as follows:

- calculate the probability to apply a data augmentation for the current training epoch using the following function:

$$Pr(use\_dg) = min([(1 + 4 * \sqrt{\sqrt{(tot\_ep)} * ep})/tot\_ep, 0.99]) \quad (1)$$

- apply the data augmentation with a probability equal to $Pr(use\_dg)$

F1-score is used as main evaluation metric; for the multi-class dataset, the F1 metric is calculated for each label (with respect to the others) in order to assess hard ans easy classes from the classification point of view.

IV. Results

From the table XIV all models performed good on Deep-Fake dataset. Concerning the Genre Dataset, both KNN and RF models recognized very well classical, pop, metal classes; while hiphop,reggae and rock can be considered hard classes. CNN model outperformed the other models(average score of 73%) proving the effectiveness and predictive power of a transfer learning approach.

V. Conclusions and Future Work

In this study traditional ML models and DL approaches have been used for both binary and multi-class classification tasks. In addition, EDA tools and unsupervised techniques have been used to gather insights from data as is the case in any business environment where it's important to explain and understand both models and features. CNN proved to be effective for all tasks, however further strategies could be used to improve both CNN and ML models. For example one can first try to explore in more details the hard classes that is classes for which a model performed bad. One way to do that is via the EDA approaches showed in this paper. Also data augmentation can be used or improved (as is the case for the CNN used for this study) on top of any ML or DL approach [23]

References

[1] G. Eason, B. Noble, and I. N. Sneddon, Deep learning Binary/Multi classification for music's brainwave entrainment beats, Nov 2023
[2] Dr. S. Vena, M. Nerisai, J. Remya, Sound Classification system using machine learning techniques, May 2020
[3] P. Ghosh, S. Mahapatra, S. Jana, A Study on Music Genre Classification using Machine Learning, Apr 2023.
[4] J. Gua, F.Xiao, Y.Liu,Anomalous sound dtection using audio representation with machine ID based contrastive learning pretraining, Apr 2023
[5] H. Rajula, G. Verlato, M. Manchia, N. Antonucci,Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnonis, Drug Development, and Treatment, Sep 2020
[6] Y. Bouchlaghem, Y. Akhiat, S. Amjad, Feature Selection: a Review and Comparative Study, "https://www.e3sconferences.org/articles/e3sconf/abs/2022/18/e3sconf_icies2022_01046/e3sconf_icies2022_01046.html",2022
[7] A.Maccagno, A.Mstropietro, U.Mazziotta A CNN Approach for Audio Classification in Construction Sites,2021
[8] Mary L. McHugh, The Chi-square test of independence, 2013
[9] M.Verleysen, D. François, The Curse of Dimensionality in Data Mining and Time Series Prediction,IWANN 2005: Computational Intelligence and Bioinspired Systems pp 758–770, 2005
[10] M. E. Tipping, C.M. Bishop, Probabilistic principal component analysis, 1999
[11] J.B. Tenenbaum, V. De Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, 2000
[12] M. Hollander, D.A. Wolfe, Nonparameteric statistical methods, 2013
[13] L. Breiman, Random Forest, 2001
[14] M. Hasan, S. Ullah, M.J. Khan, Comparative analysis of SVM, ANN AND CNN for Classifying vegetation species using hyperpspectral thermal infrared data, Jun 2019
[15] R.Keeling, N.Huberf-Fliflet, J.Zhang,F.Wei,H.Zhao,S.Ye,H.Qin, Empirical Comparison of CNN with Other Learning Algorithms for Text Classification in Legal Document Review, Jun 2019
[16] J. J. Bird, A. Lotfi, Real-Time Detection of AI-Generated Speech For DeepFake Voice Conversion, "https://arxiv.org/abs/2308.12734", Aug 2023
[17] B. L. Sturm, The GTZAN dataset, "https://arxiv.org/abs/1306.1461", JJun 2013
[18] D. Matthew, D. Saputra, L. D. Oswari, Effect of Distance Metrics in Determining K-Value in K-means clustering Using Elbow and Silhouette Method, 2019.
[19] M.Radovanovic, A.Nanopoulos, M.Ivanovic, Nearest Neighbors in High-Dimensional Data: The Emergence and INfluence of Hubs, 2009
[20] T. Wang, H. Zhang, L. Tian, The Application of high-dimensional Data Classification by Random Forest based on Hadoop Cloud Computing Platform, 2016
[21] K. He, X. Zhang, S Sen,J. Sun, Deep REsidual Learning for Image Recognition, 2015
[22] M. Andriushchenko, F. D'angelo, A. Varre, N. Flammarion, Why Do we need Weight Decay in Modern Deep Learning?, Oct 2023
[23] L.Nanni G. Maguolo, M. Paci, Data augmentation approaches for improving animal audio classification, 2019